What is retransformation bias, and how can it be corrected?

Economists frequently wish to estimate regression models using healthcare cost as the dependent variable. Health data are often strongly skewed to the right, however, making ordinary least squares unattractive. For example, the length of inpatient stays and the cost of inpatient care are often highly skewed (and kurtotic).

A common approach is to use the natural log of cost in place of raw cost. The logarithmic transformation often removes enough skewness to allow least squares models to produce unbiased results. A number of other transformations have also been studied: Poisson and negative binomial models for count data, generalized linear models (GLM), and survival (hazard) models.

Each of these involves a nonlinear transformation of the dependent variable. The resulting estimated coefficients are not directly interpretable in raw dollars. Nor can one simply reverse the transformation, as this will cause a Doing so would cause a retransformation bias (Manning 1998).

I. The case of homoskedastic errors: the smearing estimator

If the errors from the regression are homoskedastic, one can determine an appropriate retransformation through the smearing estimator (Duan 1983). Consider a regression model of the form:

$$\ln \text{cost} = \mathrm{X}\beta + \varepsilon$$

A simulation of the retransformed fitted value (cost) when $X=X_0$ is not simply:

$$\hat{\text{cost}} \neq \mathrm{e}^{(\mathrm{X}_0\beta)}$$

Although the expected value of the residual is zero, it is subject to a non-linear retransformation. The expected value of cost when $X=X_0$ is thus:

$$\mathrm{E}(\mathrm{cost}) = \mathrm{E}\left(\mathrm{e}^{(\mathrm{X}_i\beta)+\varepsilon}\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{e}^{(\mathrm{X}_i\beta)+\varepsilon_i}\right)$$
$$= \left(\mathrm{e}^{(\mathrm{X}_i\beta)}\right)\left(\frac{1}{n}\sum_{i=1}^{n}e^{\varepsilon_i}\right)$$

The smearing estimator for models with log-transformed dependent variables is the right hand factor. It is the mean of the anti-log of the residuals:

$$\frac{1}{n}\sum_{i=1}^{n}e^{\varepsilon}$$

Most regression packages allow the analyst to save the residual. To find the smearing estimator, we find the anti-log of the residuals, and then find its mean. This often yields a value between 1 and 2. The smearing estimator is then multiplied by the fitted value to correct it for retransformation bias.

Duan (1983) also describes the smearing estimator appropriate for other non-linear transformations of the dependent variable, such as the square root.

II. The case of heteroscedastic errors

Quite often, the error for a particular observation in the cost regression will depend on the level of one or more regressors. This situation, known as heteroscedasticity, precludes the use of Duan's smearing estimator.

There are several approaches to this problem. A technical explanation is beyond the scope of this FAQ response; interested readers are encouraged to read the journal articles in the References section below.

Mullahy (1998) lays out the econometric problem in detail and derives the bias of the smearing estimator when heteroscedasticity is present. Manning and Mullahy (2001) and Basu et al. (in press) describe several alternatives: ordinary least squares on the natural log of y; GLM variants (such as gamma regression with log link and Weibull regression with log link); and the Cox proportional hazards model. They conclude that no single model is best under all circumstances.

References

Basu A, Manning WG, Mullahy J. Comparing alternative models: log vs proportional hazard? *Health Economics* (in press).

Duan N. Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association* 1983;78:605-610.

Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics* 1998;17(3):283-295.

Manning WG, Mullahy J. Estimating log models: to transform or not to transform. *Journal of Health Economics* 2001;20:461-494.

Mullahy J. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics* 1998;17:247-281.

by Dr. Mark Smith & Dr. Paul Barnett

Last updated: July, 2004